Article Submission

# Kingdom-wide Analysis of Fungal Transcriptomes and tRNAs Reveals Conserved Patterns of Adaptive Evolution

Rhondene Wint[a,b], Asaf Salamov [b], Igor V. Grigoriev[b,c]

[a]Molecular and Cell Biology Unit, Quantitative and Systems Biology Program, University of California Merced, Merced, CA 95343, USA,

[b] US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA,

[c] Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720 US

Corresponding author: Igor V. Grigoriev (ivgrigoriev@lbl.gov)

Keywords: codon usage, tRNA, translation, fungi, macroevolution, machine learning

© The Author(s) 2022. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

## ABSTRACT

Protein-coding genes evolved codon usage bias due to the combined but uneven effects of adaptive and non-adaptive influences. Studies in model fungi agree on codon usage bias as an adaptation for fine-tuning gene expression levels; however, such knowledge is lacking for most other fungi. Our comparative genomics analysis of over 400 species supports codon usage and tRNAs as co-adapted for translation speed and this is most likely a realization of convergent evolution. Rather than drift, phylogenetic reconstruction inferred adaptive radiation as the best explanation for the variation of interspecific codon usage bias. Although the phylogenetic signals for individual codon and tRNAs frequencies are lower than expected by genetic drift, we found remarkable conservation of highly expressed genes being codon-optimized for translation by the most abundant tRNAs, especially by inosine-modified tRNAs. As an application, we present a sequence-to-expression neural network that uses codons to reliably predict highly expressed transcripts. The kingdom Fungi, with over a million species, includes many key players in various ecosystems and good targets for biotechnology. Collectively, our results have implications for better understanding the evolutionary success of fungi, as well as informing the biosynthetic manipulation of fungal genes.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

## Introduction

The billion-year-old kingdom Fungi, comprising at least 1.5 million species, is deeply intertwined with the diversification and maintenance of terrestrial ecosystems (Berbee and Taylor, 2017). Paleo-botanical studies owe the successful colonization of land by primitive plants – resulting in the greening of the Earth that facilitated the evolution of more complex animal forms – to their mutualistic symbioses with soil fungi (Field et al., 2015). Indeed, 90% of extant plant species still rely on mycorrhizal fungi for nutrient uptake and resistance to pathogens and abiotic stressors (Eva et al., 2018). Many fungi are also pathogens of plants, fungi and animals, and pose an emerging medical threat to humans (Janbon et al., 2019). The diversity of fungal bioproducts is leveraged in biotechnology to manufacture commercial enzymes, medicines, and even biofuel (Sepala et al., 2017). Therefore, a comprehensive understanding of the evolution of fungal genomes and traits is valuable to several applications.

Codon usage analysis is an established framework for studying the evolution of protein-coding genes. Although the genetic code assigns 61 mRNA codons to 20 amino acids, most organisms have evolved an unequal representation of synonymous codons, or codon usage bias (CUB). The widely accepted *mutation-selection-drift* theory posits that CUB is generated by evolutionary forces that imprinted non-adaptive (neutral) and adaptive mutations at the silent sites within coding sequences (Grantham,1980; Bulmer, 1991). But far from being silent, functional studies in model fungal systems such as *Neurospora crassa* and *S. cerevisiae* (baker's yeast) demonstrates how the influence of synonymous mutations percolates through all layers of gene expression, from mRNA transcription (Zhao et al., 2021), steady-state mRNA levels (Sharp et al, 1987), mRNA stability (Presnyak et al., 2015), elongation rate of protein synthesis (Tuller et al, 2010), and co-translational protein folding (Pechmann and Frydman, 2013).

Because transfer RNAs (tRNAs) universally translate codons to amino acids during protein synthesis, codon usage variation within genomes may reflect selection for balancing codon demand with tRNA supply in order to fine-tune (Hershberg and Petrov 2008). Ideally, for each of the 61 sense codons there should be 61 distinct tRNA anticodon types. However, species usually lack the full complement of tRNA types because tRNAs often engage in wobble-decoding (Marck and Grosjean, 2002). Moreover, the genomic dosage of different tRNAs take on a dynamic range, from single-copy to even hundreds of identical or near-identical copies within the same genome. This imbalance in the supply decoding tRNAs among synonymous codons establishes the selective pressure that underlies adaptive CUB in several species, including *S.cerevisiae* (Ikemura,1981). Indeed, a hallmark of adaptive

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

codon usage bias is the preference of highly abundant mRNAs for translationally optimal codons that are decoded by abundant tRNAs (Sharp et al 2010; Novoa et al, 2015). Therefore, disentangling the signals of neutral and adaptive forces on codon usage offers insight into the evolutionary processes that contribute to the fitness and biodiversity of species (Novoa et al., 2019). However, outside of model species, codon usage patterns across the fungal phylogeny remain largely uncharacterized.

Comparative studies aim to disentangle the trait variation due to shared ancestry versus adaptation. Because of common descent, phenotypic traits from closely related species are likely to violate the identically and independently distributed requirement of standard regression tests which risks an increase in type I errors. Phylogenetic comparative methods (PCMs) are regression algorithms that account for phylogenetic signal in comparative trait data (Felenstein,1985). Phylogenetic signal is the tendency of closely related species to exhibit greater similarities in traits than other species when sampled randomly from the same phylogenetic tree. The strength and direction of the phylogenetic signal are used to infer whether trait variation exhibits signs of evolution due to genetic drift, stabilizing selection, divergent or convergent evolution (Blomberg et al., 2003; Pagel et al.,1999). PCMs have been applied to interrogate macroevolutionary questions such as the evolution of fungal modes of nutrition (James, 2006), evolution of physiological and behavioral traits in primates (Kamilar and Cooper, 2010), plant-pollinator co-evolution (Smith et al,2010), trait evolution by adaptive radiation in reptiles and avians (Pichereira-Donoso et al, 2015; McEntee et al., 2018). However, the application of PCMs is rather limited in cross-species codon usage studies (Sharp, 2010; LaBella et al, 2019). Recent large-scale sequencing projects have advanced our understanding of fungal phylogeny (Grigoriev et al., 2014; Ahrendt et al., 2018), thereby broadening the scope for comparative studies.

Here, we aimed to detail the evolutionary and functional underpinnings of codon usage variation in Kingdom Fungi by analysing transcriptomic and tRNA data from over 400 representative species that are distributed across 18 taxonomic classes and 6 major phyla (Spatafora et al., 2017). Principal component analysis of codon usage frequencies effectively separated the species into respective sub-kingdoms, with the rare codons $AUA^{Ile}$ and $GGG^{Gly}$ driving the codon-specific variation. Using phylogenetic reconstruction methods, we inferred the evolutionary processes, including adaptive mechanisms, that explain change in codon usage and tRNA patterns over time. We also performed genome-level analyses to examine the relationship between codon usage, tRNA supply and gene expression levels. Phylogenetic signals of codon frequencies and genomic tRNA abundance were weaker than expected by genetic drift and phylogenetic relatedness. Yet, most genomes converged

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

toward translation bias, wherein the most abundant mRNAs are enriched with codons for major tRNAs, in contrast to the low abundant mRNAs having greater codon bias for minor tRNAs. Finally, given the prevalence of adaptive codon usage, we present a neural network, *Codon2Vec*, that directly takes the coding sequences as input to reliably predict expression (median accuracy of 83.8% ±0.05). Altogether, our results support that natural selection for the efficiency of mRNA translation is a conserved influence among fungi.

## RESULTS

### Adaptative radiation best explains the macroevolution of codon usage bias

We obtained whole transcriptomes and predicted tRNA genes from 459 species sampled from six out of the eight recognized fungal phyla (Methods). Namely, 57 species belonging to the four early-diverging phyla of *Chytridiomycota, Blastidiomycota, Zoopagomycota, Mucoromycota,* and 402 species from the two dikarya phyla *Basidiomycota and Ascomycota*. Dikarya is the more species-rich sub-kingdom comprising 98% of all fungi - but 90% of our dataset - and is characterized by a more complex sexual lifecycle (Stajich et al., 2009).

*Codon usage bias is evolutionarily correlated with the usage of GC-ending codons*

We measured the degree of codon usage bias by computing the effective number of codons, ENC, for each species (Wright, 1990). ENC ranges from 20 to 61, where 20 represents extreme bias of using only one codon per amino acid, while 61 represents uniform synonymous codon usage, that is, no bias. The mean ENC values ranged from 32.8 (high bias) to 56.9 (weak bias). To visualize the macroevolutionary pattern of codon usage bias, we applied continuous maximum likelihood ancestral state reconstruction (Revell, 2013) that projected the species ENC values onto a pruned phylogenetic tree. The ancestral reconstruction shows that the more biased genomes accumulate in the early-diverging lineages (Figure 1A) with the most codon-biased genomes occurring in *Neocallimastigomycota,* the earliest diverging class of free-living fungi (Berbee and Taylor, 2017). Also, there is more fluctuation in codon bias along the upper branches that slows down upon the divergence of *Agaricomycotina*, the largest class (~70%) in *Basidiomycota*. Similarly, species in *Ascomycota* exhibit less variation in their codon bias. Variation in the GC-content at the third codon position (GC3%) is closely linked to codon usage bias since all degenerate amino acids allow for silent G or C substitutions. The mean GC3% ranges from 10.6% to 85.1%, with a median of 57%. Overall, early-diverging fungi exhibit, on average, lower GC3% but more variability among individual values (Figure 1A).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

Next, we assessed the evolutionary relationship between codon usage bias and GC3% using phylogenetic independent contrast (PIC). PIC regression corrects for phylogenetic non-independence by using the contrasts between nodes instead of the trait values directly (Garland et al, 1992). For the entire tree, the PIC model *ENC~GC3* yielded a negative coefficient of -14.1 (adjusted $R^2$ =11.8%, p-value=$3.79e^{-13}$). Because PIC is calculated without an intercept term, the $R^2$ coefficient is the square of the Pearson's *R* correlation coefficient. Therefore, codon usage bias and GC3% are moderately correlated (Pearson's R=34.4%). Although it may be reasonable to assume the evolutionary GC3-bias is driven by the usage of G/C-ending codons, it was found that the usage of certain G/C-ending codons was negatively correlated with GC3-bias in some plants and prokaryotes (Palidwor et al, 2010). To evaluate the relationship between codon usage and GC3-bias, we computed the phylogenetic-corrected Pearson's correlation between individual codon frequencies (normalized for amino acid usage), GC3% and ENC, separately. The usage of all G/C-ending codons are positively correlated with GC3%, whereas all A/U-ending codons are anticorrelated with GC3%(Figure 1B). All G/C-ending codons negatively correlated with ENC, which means that the increase in usage of G/C-ending codons correlates with an increase in codon bias. Conversely, the usage of all A/U-ending negatively correlates with codon bias. Interestingly, we obtained different correlation values between codon bias and the normalized codon frequencies with and without phylogenetic correction. Without the correction, some A/U-codons positively correlate with codon bias, and some G/C-ending codons negatively correlate with codon bias (Figure 1C). This discrepancy suggests that codon frequencies also have phylogenetic signal.

Since the relationship between codon bias and GC3% seems inverted for the early-diverging and dikaryic lineages (Figure 1), we split the tree into the separate sub-kingdoms and re-evaluated the phylogenetic correlation between codon usage bias and GC3%. Codon bias and GC3% are evolutionarily anticorrelated in the early-diverging subtree (coefficient=21.1, $R^2$ =32.6%, p-value=$9.02e^{-06}$, number of tips=51 species). In contrast, dikarya species are positively GC3%-biased (coefficient= -30.1, $R^2$ =50.2%, p-value<$2.16e^{-16}$; number of tips=364).

*Fitting macroevolutionary models to codon usage bias*

Phenotypic variation among extant species is a confluence of shared ancestry and responses to neutral and adaptive processes. Interspecies codon usage bias is widely ascribed to neutral drift (Grantham et al., 1980). To determine the pattern of evolution that best explains codon usage bias, we fitted 4 different likelihood models of macroevolution to the ENC and GC3 values: 1) Brownian motion (drift/random walk), 2) Ornstein-Uhlenbeck

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

(fluctuating directional selection), 3) early-burst (exponential decrease in trait variation over time) and 4) delta (rate shifted Brownian motion) (Pennell et al, 2014). Notably, Brownian motion is the null hypothesis of genetic drift that models interspecies trait data as a random walk (Felestein, 1985). Based on the goodness-of-fit Akaike information criterion (AIC) scores, the early-burst model, which simulates adaptive radiation, best explained the phylogenetic variation of both codon bias and GC3% (Supp. Table 2). Macroevolution by adaptive radiation is characterized by higher rates of trait evolution early in a clade's history followed by an exponential decline through time (Simpson, 1953).

**Codon-level macroevolutionary analysis reveals codon frequencies as mostly deviate from genetic drift.**

Since variation in the frequencies of synonymous codons underlies codon usage bias, we examined the macroevolutionary trends of the individual codons. First, we quantified transcriptome-wide relative synonymous codon usage (RSCU) of the 59 degenerate codons (Sharp et al, 1986). RSCU=1 means codons are used according to neutral or uniform expectation. Importantly, RSCU normalizes codon frequencies within their amino acid class which minimizes amino acid composition effects. To characterize which codons fungi generally prefer for making proteins, we quantified the most (highest RSCU) and least (lowest RSCU) preferred codons. Overall, C-ending codons consistently had the highest transcriptomic representation across the amino acid types (Figures S1A, S1C).

To summarize variation of interspecies codon usage, we performed multivariate analysis using principal component analysis (PCA) on the 59 RSCU x 459 species matrix. The first two principal components explained 82% of the interspecies variation (Figure 2A). PC1 (78% explained variance) separated species according to differences in GC-content at the third codon position (GC3%), wherein loadings of G/C- and A/U- ending codons are equally but inversely correlated to PC1 (Figure 2B). This finding aligns with previous work that establishes variation in G+C content as the major determinant of interspecies differences in CUB (Chen et al., 2004; Novoa et al., 2019). The second principal component, PC2, (4.0% explained variance is driven by differences in individual codon frequencies, with the strongest signal due to the rare codons $GGG^{Gly}$ and $AUA^{Ile}$ (Figure 2B; S1B). Notably, PCA separated the species into their sub-kingdoms (Figure 2A).

The sub-kingdom clustering by the PCA led us to measure the extent to which phylogenetic effect (i.e., phylogenetic relatedness) underlies the choice of codon representation in the transcriptome. To this end, we computed the Blomberg's *K* statistic (Methods) of the normalized codon frequencies. Blomberg's *K* measures the strength and direction of trait

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

evolution relative to that expected under the Brownian motion model that considers the phylogenetic distance as the only predictor of trait similarity among species (Blomberg et al., 2003). All codons reported statistically significant phylogenetic signals (p-values<0.05). However, the strength and direction of evolution, even among synonymous codons, varied (Figure 2C). 37/59 codons exhibited low phylogenetic signal (K<1), suggesting variation due to convergent evolution (Revell et al., 2008; Kamilar and Cooper, 2013). 8 out of the 59 codons followed the expected Brownian process (K=1) of genetic drift. 14 out of the 59 codons exhibited high phylogenetic signal (K>1) indicative of either stabilizing selection or low rates of evolution (Blomberg et al, 2003). We also fitted different models of macroevolution to the individual codon frequencies. Like genomic CUB, adaptive radiation was the best fitting model for all the 59 degenerate codons (Supp. Data). Taken together, these findings highlight that individual codons follow different modes of evolution. Importantly, the frequencies of 51 out of 59 codons are not fully explained by phylogenetic relatedness that is expected under genetic drift.

**Identification of phylogenetically rare tRNAs and strong evolutionary preference for Inosine34-modified tRNAs**

Considering that the frequencies of most codons deviated from genetic drift, the next logical step was to analyze the tRNA gene sets since codon usage is widely believed to co-evolve with tRNA supply in several species (Sharp et al., 2010). Both the number of distinct tRNA anticodon types and total tRNA genes (tRNAome) vary widely across the 459 genomes under study. The median number of distinct anticodon types is 44, ranging from a maximum of 58 in *Ascobolus immersus* and a minimum of 18 in *Sporobolomyces linderae* (Figure S2A). Like all previously studied genomes, no species in our dataset possessed the full theoretical complement of 61 tRNA anticodon families (Marck and Grosjean, 2002). Interestingly, we identified 11 species possessing less than 30 anticodons, which is the theoretical minimum for decoding the standard genetic code (Marck and Grosjean, 2002). The median tRNAome is 144 genes, with a minimum of 24 and a maximum of 4547. *Caecomyces churrovis, Ascobolus immerses* and *Melampsora allii-populina* possessed unusually large tRNAomes of 4547, 3481 and 2216 genes, respectively (Figure S2B).

Next, we measured the phylogenetic signal of the copy number for tRNAs that are cognate to the 59 degenerate sense codons. Like most codons, tRNAs also exhibited phylogenetic signal that is lower than expected by drift, with K ranging from 0.02 to 0.70. These weak phylogenetic signals are consistent with tRNA gene dosage as evolutionarily labile (Velandia-Huerto et al., 2016). 44 out of 59 sense tRNAs yielded statistically significant phylogenetic signal (p<0.05). The lack of phylogenetic signal (p>0.05) in the remaining 15

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

tRNAs implies that they either evolved completely independent of phylogeny or are mostly absent in the fungal genomes since entire tRNA families are known to be extinct in certain clades (Rak et al., 2018). To this end, we identified 19 tRNA anticodon types that rarely occur among the fungal genomes, three of which are nonsense suppressors ($tRNA^{Sup}$ (UUA), $tRNA^{Sup}$ (CUA), $tRNA^{SeC}$ (UCA)) (Figure 3A). 14 out of the 16 rare sense tRNAs overlapped with the 15 tRNAs that lacked phylogenetic signal (Figure 3B). Only $tRNA^{Ile}$ (UAU) is prevalent yet lacking a phylogenetic signal. In other words, close relatives are no similar in their genomic copies of $tRNA^{Ile}$ (UAU) than if they were randomly placed on the tree. This finding may be explained by highly accelerated birth-death evolution or anticodon shifts of $tRNA^{Ile}$ (UAU) along the phylogeny (Velandia-Huerto et al., 2016).

*Detection of selenocysteine-tRNAs in dikarya genomes*

Here, we would like to report the detection of selenocysteine tRNA (SeC-tRNA). At the time of this finding, tRNAs corresponding to the 21$^{st}$ amino-acid selenocysteine were considered absent in all fungi (Lobanov et al., 2007) until Mariotti et al uncovered the presence of $tRNA^{SeC}$ (UCA) in nine early-diverging fungi (Mariotti et al., 2019). However, all three of our Sec-tRNA positive fungi – *Rhodocollybia butyracea, Sugiyamaella americana,* and *Lollipopaia minuta* – are dikarya from *Basidiomycota* and *Ascomycota* phyla (Supplemental Table 3). We identified the presence of $tRNA^{SeC}$ (UCA) in these three genomes based on overlapping results from at least one of the general-purpose tRNA gene finders, tRNAscanSE2.0 (Chan and Lowe, 2106) or aragorn1.2.38 (Laslett and Canback, 2004), and the specialized $tRNA^{SeC}$ gene finder Secmarker (Santesmasses et al., 2017). As a negative control, we repeated the analysis on the well-studied fungal genomes of *S.cerevisiae* and *N.crassa*. Even with the unrealistically relaxed parameters, $tRNA^{SeC}$ (UCA) was not detected in either genome.

Next, we examined the prevalence of inosine-modified tRNAs. Adenosine-to-inosine (6-deaminated adenosine) conversion is the most common post-transcriptional editing in eukaryotic RNAs (Nishikura, 2016). In eukaryotes, A34-to-I34 conversion is restricted among the eight tRNA types: $tRNA^{Thr}$(AGU), $tRNA^{Ile}$(AAU), $tRNA^{Pro}$(AGG), $tRNA^{Arg}$(ACG), $tRNA^{Leu}$(AAG), $tRNA^{Ala}$ (AGC), $tRNA^{Val}$(AAC), and $tRNA^{Ser}$(AGA). Inosine-34 tRNAs (INN) decode both NNC and NNU codons in eukaryotes (Rafels-Ybern et al, 2017). Although both INN and GNN tRNAs decode C-ending codons, the I:C anticodon-codon bond is known to be less stable than G:C bond (Hoernes et al; 2018). Yet, we found that for the amino acids that are recognized by isoacceptor pairs of GNN and a putatively inosinylated ANN, the GNN iso-acceptor is mostly absent within the genomes (Figure 3C). For example, $tRNA^{Leu}$ (GAG) is the rarest tRNA being predicted in only 1/459 species (Figure 3A), yet its Watson-Crick

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

cognate codon CUC is the commonly most preferred for encoding leucine (347/459 species; Supp. Figure 2A). Likewise, the usage of AUC codon is frequently the most preferred for isoleucine, yet its cognate $tRNA^{Ile}$ (GAU) is rarely present (16/459 species). According to wobble rules, both Leu-CUC and Ile-AUC codons are decoded by inosine-modified $tRNA^{Leu}$ (AAG) and $tRNA^{Ile}$ (AAU), respectively. In contrast, when the ANN tRNA is not a target of inosinylation, the GNN iso-acceptor is far more prevalent (Figure 3D). As previously mentioned, genomes in our dataset are mostly biased for NNC codons (S2A) so this finding suggests that inosine-modified tRNAs are positively selected for in fungi. To summarize, phylogenetic comparative analyses revealed that the interspecies variation of codon usage bias and individual codon frequencies do not support genetic drift as the dominant mode of evolution.

**Signatures of neutral and adaptive evolution on intra-genomic codon usage bias**

Having analyzed codon usage patterns at the macroevolutionary scale, we next sought to disentangle the signatures of adaptive and neutral evolution on within-genome codon usage bias. At the organismal level, codon usage bias is a composite of drift, neutral and adaptive mutational bias (dos Reis, 2009). To determine whether codon usage bias is driven solely by GC-compositional mutational bias in each species, we compared the empirical effective number of codons (ENC) of all coding sequences to their theoretical ENC that is expected under the neutral-mutational model. The neutral-mutational model is the null hypothesis that selection pressure does not act on the synonymous third codon position sites; rather codon bias is only a function of GC3% (Wright 1990). The codon usage bias of 458 out of 459 species deviated significantly from neutral expectation (paired Wilcoxon signed-rank test p-values<0.05; Figure 4A). Next, we assessed each species' fit to neutral expectation by computing the $R^2$ between empirical and theoretical ENC (LaBella et al, 2019) of all coding sequences within each genome (Figure 4B). The $R^2$ values ranged from 0.0001 to 0.88. 70 genomes, mostly dikarya, reported an $R^2$ value of at least 0.75 ('Very Strong'), which indicates that their codon usage bias is largely influenced by neutral mutational bias. Notably, early-diverging species make up 12% of the dataset, but 28% of the genomes with 'Weak' neutral mutational bias.

*Fungal transcriptome-wide codon usage and tRNA copy number are positively correlated*

Another signature of natural selection is the correlation between codon frequencies and the supply of cognate tRNAs (Sharp et al, 2010). To explore this, we computed the Spearman's rank coefficient between transcriptome-wide relative codon frequencies (RSCU) and tRNA gene copy number. Codon frequency and tRNA copy number were significantly correlated in

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

312 of the 459 species (p-values<0.05), all of which yielded positive correlations (mean Spearman's ρ=0.49; Fig 4C). Those species with non-significant correlation generally possessed single-copy tRNA gene sets. Additionally, the most overrepresented codons – i.e. highest RSCU – are cognates of tRNAs with a higher copy number (mean tRNA gene copy number of 5.2) compared to the most underrepresented codons (mean tRNA gene copy number of 1.4, one-sided paired Wilcoxon signed ranked test p-value = $5.50e^{-53}$; Figure 4D).

In summary, we showed that variation in the genome-level codon usage bias is influenced by both neutral (GC3-composition) and adaptive mutational bias (cognate tRNAs).

**Differential adaptation to the tRNA pool underlies expression-linked codon usage bias**

Here, we explored the functional implications of adaptive codon usage bias by analyzing the contribution of gene expression to codon preferences. Because we have RNAseq data for most species in our dataset (420/459) we could empirically investigate expression-linked codon usage bias. To this end, we filtered the top 10% ('high') and bottom 10% ('low') of expressed coding sequences as the working dataset for each species since directional selection acts at the extremes. We observed that for most species, PCA on the 59-dimensional matrix of codon frequencies (RSCU) separated the genes according to expression level (Figure 5A; Supp Data). The trend suggests that gene expression level is a driver of codon usage patterns.

*Not an artifact: PCA arch of high expressed genes caused by strong deviation from neutral compositional bias*

Intriguingly, the pattern of only the high expressed genes clustering as an arch in *Z. heterogamus* also appeared in 26 other species (left panel Figure 5A; Supp Data). Guttman or arch effect in dimensionality reduction techniques, such as PCA or correspondence analysis, is observed when the first two transformed axes are curvilinear because the structure of the data is dominated by a single latent variable that gradually shifts from one extreme to the another, i.e., the data points lie on a gradient (Diaconis et al 2008; Ngyuen et al, 2019). Given that codon usage is influenced by both directional neutral and selection pressures, we hypothesized that the latent gradient underlying the high expressed genes represents the shift in influence of neutral to selection pressures. To explore this in *Z. heterogamus*, we generated an ENC-GC3 neutrality plot in which the standard curve represents the expected relationship between ENC and GC3% when codon usage is solely explained by neutral compositional bias (Wright 1990). The neutrality plot confirmed our hypothesis as the genes are more C3-biased with increasing distance from the neutral curve

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

(Figures 5B). However, the codon usage of low expressed gene set did not exhibit such marked deviation from neutral expectation (Figures 5C). Indeed, the latent gradient responsible for the arch is defined by diametric usage of C-ending (Figure 5D) and A-ending(S3A-C) codons. We identified a similar pattern for 18 of the remaining 26 'arch' species in which their high expressed genes also lie on a C3% gradient when projected onto the first two principal components (Figure S3G), and deviating from the expected neutral compositional bias(Figure S3H). This led us to revisit the compositional bias analysis performed in the previous section, where we found that 23 out of these 27 species, including *Z.heterogamus,* fell in the 'Weak' category (Figure 4B, S3D). Additionally, 23/27 of them are AU3-biased (mean GC3 <50%), and 21 of them are early-diverging fungi (Figure S3D). Together, these results suggest that selection is particularly stronger in the highly expressed genes of these species.

*High expressed genes preferentially use codons decoded by major tRNAs but avoid codons decoded by minor tRNAs*

We then asked if the divergent codon preference between the high and low expressed genes is related to adaptation to translation efficiency (Duret, 2000). To this end, we measured the fraction of preferred and non-preferred codons that are decoded by major and minor tRNAs per species. Preferred codons are significantly enriched (higher RSCU) in the high expression gene set, whereas non-preferred codons are enriched (higher RSCU) in the low expression set (Benjamini-Hochberg adjusted p-values <0.05) (Yannai et al., 2018). We observed that C-ending codons are mostly preferred by high expressed genes compared to A-ending codons in low expressed genes (Figure S3E). Major and minor tRNAs have the highest and lowest copy numbers, respectively, within an amino acid class. Overall, highly expressed genes preferentially use codons that are decoded by major tRNAs, which is indicative of selection for rapidly translated codons (Figure 5E). On average, 43% of preferred codons are recognized by major tRNAs compared to 24% by minor tRNAs (Figure 5E; paired Wilcoxon signed-rank test p-value = $4.26e^{-64}$ ). Conversely, non-preferred codons better matched minor tRNAs (mean fraction = 34%) than major tRNAs (Figure 5F; mean fraction = 18%; Wilcoxon p-value=$1.53\ e^{-62}$).

Because we identified the widespread preference of inosine-modified tRNAs, we extended our analysis to account for inosine-34 wobble decoding. This resulted in a marked increase in the mean fraction of preferred decoded by major tRNAs from 43% to 65% (Wilcoxon p-value = $7.47e^{-68}$), but the mean fraction of preferred codons matching minor tRNAs remained the same. For example, the match rate in *Z.heterogamus* rose from 56% to 81%. In 84% of the species, at least 50% of their preferred codons are cognates of major tRNAs when

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

inosine34 decoding is considered (Figure 5E). However, the inclusion of I34 modification did not substantially alter the fraction of non-preferred codons decoded by minor tRNAs or major tRNAs (18%; Wilcoxon p-value = $4.34e^{-60}$) (Figure 5F). Therefore, the codon bias in high expressed genes can be partially explained by selective usage of inosine34 decoded codons. These results align with experiments in mammalian and bacterial systems that demonstrated improved agreement between codon usage and tRNA abundance when I34 modification is accounted for and that transcripts with codon compositions that matched I34 tRNAs were more efficiently translated (Novoa et al., 2012).

*Prevalence of codon optimization for translation in high expressed genes shows signs of convergent evolution*

To quantify the association between expression-linked codon bias and adaptation to the tRNA supply in a genome, we derived the translation bias score (TBS) (Methods). Formally, TBS is the difference between the fractions of preferred codons and non-preferred codons for major tRNAs normalized by their sum. A TBS of +1 indicates that the codon bias of highly expressed genes confers them exclusive access to the most abundant tRNAs in the cellular pool compared to the lowest expressed genes; whereas a TBS of 0 means that there is no competition for major tRNAs between the high and lowest expressed gene sets. Among the 420 species analyzed, the mean TBS is +0.39 when restricted to Watson-Crick pairing, and a mean of +0.53 when I34 wobble is considered (Figure 5G). That is, on average, 53% more of the major tRNAs are decoding high expressed genes than the low expressed genes, which we interpret as selection for translation speed. However, there are a few species that possess negative TBS meaning their low expressed genes are more codon biased for major tRNAs.

We wondered if the positive skewness of translation bias scores (TBS) toward was mainly a consequence of phylogenetic relatedness is, since species richness is unevenly distributed along our fungal tree. To measure the strength of phylogenetic effect of the TBS values, we computed Blomberg's K statistic which yielded K=0.12 for Watson-Crick pairing and K=0.18 for inosine-34 wobble decoding (both p-values = 0.01). These low *K* values indicate that distantly related species have more similar translation bias scores than expected by phylogenetic distance, a pattern often attributed to convergent evolution (Revell et al, 2008). As a complementary approach, testing of different macroevolutionary models supports the Ornstein-Uhlenbeck process of fluctuating directional selection as the best fit for translation bias (Figure 5H). Additionally, the ancestral state reconstruction shows that similarly high translation bias is distributed across multiple and distant lineages (Figure S3F). Therefore, both phylogenetic methods agree on adaptive codon usage bias - at least in the context of

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

gene expression - as a realization of convergent evolution. Altogether, the concordance between tRNA supply and expression-linked codon preferences supports selection on codon usage for translation accuracy and speed in fungi.

### *Codon2Vec* neural network for predicting the expression of coding sequences

Building predictive models for gene expression remains a pertinent challenge in genomics. Inspired by natural language processing models (Mikolov et al., 2013), we implemented a 3-layer neural network - Codon2Vec - that predicts expression class ('high' or 'low') directly from input coding sequences (Figure 6A; Methods). A neural network is a supervised algorithm that can model complex non-linear patterns that underlie the data. The first layer of Codon2Vec performs featurization of input sequences by mapping each codon type to a real-valued vector or 'embeddings' in Euclidean space. The codon embeddings are adjusted during model training to minimize the error between the predicted and ground truth labels.

To achieve a balanced dataset, we trained Codon2vec on the coding sequences from the top and bottom 10% expression. The training data was split into 70:20:10 for training: validation: test sets. Model selection was determined based on the training and validation sets. Final predictive performance was reported on the hold-out test set using the following metrics: misclassification error, area-under-the-receiver-operator-characteristic curve (AUC-ROC), sensitivity, and specificity (Methods). An AUC-ROC of 0.5 indicates that a model failed at learning and instead makes random predictions. When applied separately to 300 different species, Codon2Vec achieved a high median AUC-ROC score of 83.8% (Figure 6C). Randomizing the association between the input and class labels ablated Codon2Vec's discriminative power and drove the AUC-ROC to 0.5 or random predictions (Figures 6C; S4D).

We hypothesized that the model's decision boundary is the differential codon bias that exists between the sequences in the high expression and low expression classes. To this end, we computed the difference between the mean ENC of the expression classes (DOM-ENC) and measured the Spearman's rank correlation between the species-specific AUC-ROC scores and the DOM-ENCs, but there was no significant correlation (R=0.1, p-value = 0.074). Since codons are the features, we repeated the procedure using the frequency of optimal codon (Ikemura 1981) (DOM-FOP). This resulted in a significant and positive correlation (Spearman's rank coefficient R=0.45, p-value= $1.05e^{-16}$) (Figure 6D). We interpret this as the model performing better on genomes that have a wider margin of optimal codon content between the high and low expressed genes. As a sanity check to see if the length of coding sequences was a confounding variable, we found no significant correlation (R=0.1, p-

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

value=0.0755). Remarkably, Codon2Vec learned the intrinsic differences in optimal codon content between high and low expressed genes even though we did not explicitly provide this sequence property.

## Discussion

Much of our understanding of codon usage is inferred from collating findings across single-species analyses. To better detail the evolutionary mechanisms that have shaped codon usage patterns through time, we employed a phylogenetic comparative approach to analyze hundreds of representative species that span the major phyla of Kingdom Fungi. We showed neglecting phylogenetic effect can lead to different conclusions about the influence of individual codons on the degree of codon bias (Figure 1C). Our macroevolutionary analyses support, contrary to the widely held neutral-drift hypothesis, adaptive mechanisms as the driver of interspecies codon usage patterns in fungi. Fitting of various likelihood models of trait evolution to our 452-taxa phylogenetic tree showed that variation in codon usage bias and GC3% best fit the pattern generated by adaptive radiation. Additionally, the phylogenetic effect on most codon frequencies was found to be stronger or weaker than expected by random drift, a sign that is usually interpreted as stabilizing selection or convergent evolution, respectively (Losos 2011, Revel et al 2008). Adaptive codon usage was also evident at the genome-level analyses. Gene expression level and codon usage were broadly correlated as principal component analysis separated the highest and lowest expressed genes based on their codon usage patterns. In some fungi, primarily early-diverging, the deviation of high expressed genes from compositional bias was strong enough to dominate the signal captured by both principal components resulting in a Guttman effect. Since differential codon bias could arise by a neutral mechanism such as GC-biased gene conversion (Marais, 2003), we demonstrated how this prevalent pattern of expression-linked codon bias reflected differences in translation efficiency. Broadly, the high expressed genes preferentially used codons matching the most abundant tRNAs; whereas the low expressed genes were more biased for codons read by the least abundant tRNAs. Moreover, the pervasive trend of codon optimization of high expressed genes for translation efficiency, which we quantified using our translation bias scores, infers convergent evolution as the phylogenetic effect was significantly weaker than expected by Brownian motion trait evolution. Altogether, these findings are consistent with the influence of natural selection on codon usage to promote translation efficiency. Our results on the prevalence of adaptive codon usage bias in fungi are consistent with the recent sub-phylum wide codon usage analysis of *Sacchoromycotina* budding yeasts (LaBella et al,2019).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

*Macroevolutionary analyses of codon usage reveal the influence of adaptive mechanisms*

Although claims about codon usage are usually based on single-species analyses, we believed that inferences about the mode and tempo of codon usage macroevolution would further elucidate the adaptive significance of codon usage patterns. Principally, we found that the tempo of codon usage bias (CUB), and the evolutionarily correlated GC3%, in our 452-taxa tree best follow the pattern of adaptive radiation. Other fungal phylogenomic studies, primarily in mushroom-forming (*Agaricomycetes*) lineages, have also reported evidence of adaptative radiation for certain morphological traits (Varga et al., 2018; Gaya et al., 2015; Nagy et al., 2012). Various hypotheses exist for the intrinsic and ecological drivers of fungal radiations, including the evolution of complex fruiting bodies (Varga et al., 2018), transition to mutualism (Sanchez-Garcia and Matheny, 2017), and defense mechanisms (Gaya et al., 2015; Nagy et al., 2012). Previous studies have linked CUB to ecological specialization (Roller et al., 2013; Botzman et al., 2011). Badet et al uncovered that generalist parasitic fungi are more codon biased than non-parasitic fungi (Badet et al., 2017). Our finding raises the question of what are the ecological opportunities that underlie the macroevolution of codon usage bias (CUB). Visual inspection of our ancestral reconstruction shows that the decreased in the variability of CUB coincides with the divergence within *Basidiomycota*. *Basidiomycota* (club fungi) comprises about one-third of all fungi (Stajich et al., 2009). Saprophytic *Agaricomycotina* accounts for two-thirds of basidiomycetic fungi, whereas the two minor but earlier diverged classes - *Puccinomycotina* and *Ustilaginomycotina* – are mostly plant parasites (Mao et al, 2019). Relatedly, ancestral reconstruction of fungal nutritional modes showed that parasitism is non-randomly distributed along the tree and more prevalent in earlier-diverged lineages (James et al., 2006). Taken together, the evolution of CUB in fungi may be connected to lifestyle adaptation. We believe that a deeper study of the macroevolutionary relationship between CUB and the various ecological specialization in fungi is needed.

*Selection for translation efficiency may explain convergent codon usage in fungi*

Macroevolutionary analyses revealed that variation in synonymous codons is mostly convergent. The normalized frequencies of 37/59 codons yielded significantly low Blomberg's K values, indicating distantly related lineages are more similar in their codon choices than expected by phylogenetic relatedness, i.e., a sign of convergence (Kamilar and Cooper, 2013). Causes of convergent evolution are generally attributed to either shared constraints (molecular/ genetic/ physiological/ ecological, etc.) that limit or bias the production of phenotypic variation, or, to a lesser extent, random chance (Losos, 2011a).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

Here, we reason that the macroevolutionary convergence of codon usage frequencies reflects the shared constraints imposed by neutral - for example, GC-compositional bias - and adaptive pressures, for example, selection for balancing codon representation with tRNA supply (Figures 4B, 4C, 5E). Moreover, the maximum likelihood best fits the translation bias scores to the Ornstein-Uhlenbeck process of optima-directed trait evolution (Butler et al, 2004), lends further evidence that adaptative mechanisms have influenced codon usage patterns over time. That is, the convergence of highly expressed genes being codon-biased for the most abundant tRNAs compared to low expressed genes suggests that efficient protein synthesis is one of the selective optima that has constrained fungal codon usage over time. This assertion is consistent with genome-engineering experiments that first demonstrated how codon optimization in highly expressed genes exerts global effects on cellular fitness by promoting rapid turnover of free ribosomes enabling translation initiation on other transcripts (Frumkin et al, 2018).

Selection for translation efficiency is expected to favour those codons with better anticodon-codon pairing kinetics (Higgs and Petrov, 2008). Possibly, the weaker I:C anticodon-codon bond (Hoernes et al., 2018) promotes faster dissociation of the discharged tRNA from the ribosomal E-site, leading to less ribosome pausing and more available free ribosomes. This model may explain the conserved preference for inosine-modified tRNAs (INN) over GNN isoacceptors (Figure 3C), especially the general bias for $tRNA^{INN}$ decoded codons – primarily C-ending – observed in high expressed genes (Figure 5E). In light of this, a component of the general GC3-bias among fungi is likely a C3-bias due to selection for decoding by inosine-modified tRNAs.

Here, we highlight the limitations of our study and potential areas for improvement. We assumed that tRNA concentration scale with tRNA copy number which is the general case in unicellular organisms, for example, chromatin profiling in *S.cerevisiae* revealed all tRNA genes as transcriptionally active (Harismendy et al, 2003). Like all statistical models, inference from phylogenetic comparative methods (PCM) is constrained by assumptions and uncertainty. The main assumptions of PCMs are: 1) the phylogenetic tree is accurate 2) all the extant taxa are represented 3) there is measurement error in the trait data. Our 452-taxa tree does not preclude biased inference due to uneven taxon sampling since we used a limited number of representative species per clade. While we are using the best available molecular tree, given the vastness of the kingdom Fungi and ongoing sequencing campaigns we foresee updates in the fungal phylogeny (Ahrendt et al, 2018). Lastly, various evolutionary processes may give rise to the same phylogenetic pattern and current

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

macroevolutionary models may be limited in their capability to capture more complex patterns of trait evolution (Losos et al, 2011b).

In a minority of species, the low expressed genes were more codon biased for the major tRNAs (Figure 5G). This rather counterintuitive finding joins two previous works that challenge the default view that selection is reserved for codon usage of highly expressed genes (Zhou et al., 2009; Yannai et al, 2018). Codon usage in low expressed CDS may be influenced by selection for mRNA structure, mRNA stability to support sufficient protein production, or co-translational protein folding of structural sites that are sensitive to translation speed or accuracy (Zouridis et al, 2008; Zhou et al., 2009). Other than translation efficiency, the co-variation between codon usage bias and gene expression levels may reflect selection for mRNA stability as certain codons mitigate ribosomal stalling, as observed in *S.cerevisiae* (Presynak et al, 2015), or linked to transcriptional efficiency as seen in *N. crassa (*Zhou et al 2016; Zhao et al, 2021).

We also identified fungi possessing less than the theoretical minimum of 30 tRNA anticodons required for standard mRNA translation (Marck and Grosjean et al. 2002). Interestingly, 7 out of these 11 species are mutualistic symbionts - *Sporobolomyces linderae, Cenococcum geophilum, Meliniomyces bicolor, Neocallimastix californiae* – or pathogens/parasites - *Teratosphaeria nubilosa, Mixia osmundae, Elsinoe ampelina*. Perhaps their reduced tRNA gene set reflect lifestyle adaptations such as selection for rapid DNA replication, or importing necessary tRNA molecules from the host - a rare mechanism for eukaryotes that was first observed in plasmodium parasites (Bour et al., 2016). This mechanism may specifically explain how *Mixia osmundae* maintains survival as a biotrophic intracellular parasite in plants (Toome et al, 2014). Also, these minimalist fungi may also employ promiscuous super-wobbling decoding, as observed in plastomes of vascular plants (Rogalski et al., 2008). Therefore, these minimalist symbionts would make ideal candidates for studying non-standard translation of the genetic code and the co-evolution of decoding strategies within a eukaryotic host-symbiont pair.

*Codon2Vec: Addition of a sequence-to-expression model to the functional codon usage toolkit*

We believe the value of Codon2Vec is twofold. Because the model is trained on whole coding sequences, it learns a more biologically meaningful representation of the codon usage patterns. Codon order, such as codon-pair bias, has been shown to also contribute to protein yield of highly expressed genes (Cannarozzi et al., 2010; Gamble et al, 2016). But standard codon-based approaches are limited in capturing effects due to codon frequency.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

Because the model algorithm represents codons as vectors ('embeddings') in Euclidean space, in principle, contextually related codons are projected close together embedding space (Mikolov et al, 2013). Secondly, unlike standard approaches, Codon2Vec is not restricted to a pre-defined reference set of genes. Moreover, Codon2Vec bypasses the need for artisanal feature selection since it extracts information directly from sequences and expression data, and the function that maps codons to real-valued vectors is also learned during training. Although neural networks are regarded as decision ' black-boxes ', we showed that the model is at least learning to classify coding sequences based on differences in codon optimality. However, neural networks can learn complex functions there may be other sequence/codon usage properties that it may have learned. Embedding neural networks have also been used for other biological applications, like predicting chemical physicochemical properties from protein sequences (Yang et al., 2018) and gene annotation (Du et al., 2018). Once trained on the host's gene expression data, Codon2Vec can then serve as an oracle to guide the codon optimization of exogenous genes. A nice follow-up would be to experimentally validate Codon2Vec's predictions in optimizing heterologous gene expression system.

In conclusion, by combining genomics and macroevolutionary analyses we characterized the significance of and prevalence of adaptive processes in shaping fungal protein-coding genes. In the age of 'big genomics data', it would be interesting to see if similar macroevolutionary modes and mechanisms explain interspecific codon usage variation in other clades.

## FIGURE LEGEND

**Figure 1: Inferring the tempo and mode of the evolution of codon usage bias and GC3% in fungi**

Ancestral reconstruction of the codon usage bias, measured by the mean effective number of codons (ENC), and GC3-content projected onto a pruned fungal phylogenetic tree (number of tips =417 species). Color gradient represents the trait values for species at the tips and estimated trait values for internal nodes. Species with higher codon usage bias, i.e., lower ENC, and low GC3% primarily accumulate in the early-diverging lineages. Size of tree obscures tip labels but greater details are available in supplementary data. Photo credits: https://mycocosm.jgi.doe.gov/mycocosm/

**B:** Cluster heatmap showing phylogenetic corrected Pearson's R correlation between normalized codon usage, GC3-content (GC3%) and ENC which is the inverse of codon bias. G/C-ending codons are all positively correlated with GC3% and codon bias, whereas as A/U-ending codons are all negatively correlated with GC3% and codon bias.

**C:** Scatterplot showing Pearson's R correlation coefficients between individual codon frequencies and codon bias (ENC) with and without correcting for phylogenetic signal.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

**Figure 2: Phylogenetic analysis of transcriptome-wide codon usage frequencies**

**A:** Principal component analysis on RSCU matrix clusters species into the dikarya and early-diverging sub-kingdoms, primarily along the axis of PC2. Each dot is a species whose color and shape represent its sub-kingdom.

**B:** PCA biplot of the loadings showing the contribution of each codon to PC1 and PC2. Codons are colored based on the G/C or A/U composition of the third base. A/U and G/U-ending codons equally but inversely contribute to the PC1 score of a species. On the other hand, differences in PC2 species scores primarily correlate with the usage of the rare codons, $GGG^{Gly}$ and $AUA^{Ile}$ (see also S1D).

**C:** Stripplot shows the variation in the phylogenetic signal of the frequencies of degenerate codons. 51/59 codons reported Blomberg's K not equal to 1 suggesting they evolved at a rate less than or greater than the rate expected by genetic drift modeled by Brownian motion (BM). All p-values are statistically significant (<0.05).

**Figure 3: Analysis of tRNA gene composition across the phylogeny**

**A:** Prevalence of each tRNA anticodon type across the 459 fungal genomes based on tRNAscanSE2.0 predictions. Low-quality tRNA and pseudogenes with a covariance score below 50.0 are not included. Rare tRNAs are highlighted in red.

**B:** Overlap between tRNAs rarely present in fungal genomes and tRNAs with non-significant phylogenetic signal (p-value >0.05 for Blomberg's K). $tRNA^{Ile}$(UAU) is the only phylogenetically non-significant tRNA that is not rare. This suggests that this tRNA gene's evolution is de-coupled from phylogeny.

**C, D:** Evolutionary bias for the inosine-34 modification**.** For amino acids decoded by both ANN and GNN tRNAs, when the first anticodon position is a target of A-to-I editing, the INN tRNAs are more prevalent while the GNN isoacceptor is rare. **D**: However, if the ANN tRNA is not a target of A-to-I editing then the GNN isoacceptor is more prevalent and the ANN is rare.

**Figure 4: Signatures of mutational bias and natural selection on within-genome codon usage bias**

**A:** Deviation of genomic ENC values from Wright's neutral mutational model**.** The outer histogram shows the distribution of the mean difference between the empirical ENC and theoretical ENC that is expected by GC3-compositional bias measured in each of the 459 species. Inset displays the Wilcoxon signed-rank p-values (log base 10) that measures the significance of deviation. 458/459 species reported significant p-values (right of red dashed line). Both y-axes represent the number of species.

**B.** Variation in the influence of neutral pressures on species' codon usage bias. Swarmplot of species' $R^2$ values (n points =459 species) that measures the fit between empirical codon usage bias ('ENC_obs') and theoretical codon usage bias ('ENC_theo') expected solely due to GC3-compositional bias, grouped by 'Very Strong' ($R^2$ >=0.75), 'Strong' (0.75> $R^2$ >=0.5), 'Moderate' (0.5 > $R^2$ >=0.25) and 'Weak' ($R^2$<0.25).

**C:** Codon frequency correlates with tRNA copy number. Histogram shows the distribution of Spearman $\rho$ correlation coefficient between RSCU and cognate tRNA gene copy number. 312/459 species reported statistically significant p-values (p<0.05). The black line is the mean correlation coefficient of all species.

**D:** Histogram showing differences in the tRNA copy numbers for the codons with highest and lowest representation (RSCU) in the transcriptome over all the 459 species **Inset**: Distribution of the log base 10 copy numbers of tRNAs (y-axis) for the codons with highest (red) and least (blue) transcriptome-wide RSCU. The most frequently used codons are decoded by tRNAs with higher copy numbers, whereas the least frequent codons are decoded low copy-number tRNAs.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

**Figure 5: Expression-linked codon usage bias correlates with tRNA supply**

**A:** High and low expressed genes exhibit different codon usage patterns. Example of PCA applied to the codon usage 59-dimensional RSCU matrix of the top ('high') and bottom ('low') 10% expressed genes. Each dot is a gene. The right panel is the more common cluster pattern. The left panel depicts Guttman ('arch') effect in the high expressed genes of *Z.heterogamus*.

**B-D:** ENC-GC3 plot of *Z.heterogamus* genes elucidates source of PCA arch effect. The solid red line in scatterplots B and C represents the expected curve when codon usage bias is only affected by neutral mutation pressure. **B)** Codon usage bias of high expressed genes deviate strongly from neutral mutation pressure while becoming more C3-biased but **C)** low expressed genes better fit expected neutrality. **D)** Arch effect captured the variation in C3% due to neutral and selection pressures.

**E and F:** Highly expressed genes are biased for translationally optimal codons **E)** Fraction of preferred codons (significantly enriched in the top 10% expressed genes) that are decoded by major (most abundant) and minor (least abundant) tRNA isoacceptor per genome (n=420 species) with and without the inclusion of inosine-34 modification **F)** Paired comparison of the fraction of non-preferred codons (significantly enriched in the bottom 10% expressed genes) recognized by major and minor tRNAs.

**G:** Distribution of the translation bias scores (TBS) across all 420 species. TBS is the difference between the fraction codons enriched in the 10% highest versus lowest expressed transcripts normalized by their sum. The positively skewed distribution indicates that the highest expressed transcripts are generally codon-biased for rapid translation.
**H:** Akaike information criterion (AIC) goodness-of-fit evaluation of the various maximum likelihood macroevolutionary models fitted to translation bias scores (n species tips=384). 'BM'=Brownian motion, 'OU'= Orhnstein-Uhlenbeck. The OU model of fluctuation directional selection yielded the lowest AIC, and therefore the best fit model.

**Figures 6: Neural network uses codons as features to predict gene expression.**

**A:** Schema of Codon2Vec. Codon2vec is a fully connected multi-layer neural network that uses an embedding layer to transform codons in the input coding sequences to a real-valued vector. The final output of the model is a vector of probabilities for each gene expression class (i.e., "high" or "low"). Detailed description in Methods.
**B**: Codon2Vec's performance on a single species. Model performance is evaluated on the hold-out test sets based on the area under the curve of the receiver-operating-characteristics curve (AUC-ROC). An AUC score of 0.5 (dashed line) represents random predictions.
**C:** Model generalizability: Codon2Vec achieves high predictive performance on 300 different species. However, shuffling of ground truth labels independent of input sequences ablated Codon2Vec's ability to learn meaningful associations.
**D:** Codon2Vec's prediction accuracy (AUC-ROC) positively and significantly correlates with differential usage of optimal codons between high and low expressed genes. The frequency of optimal codons (FOP) is another standard CUB metric (Ikemura 1981; Methods).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

## MATERIAL AND METHODS

### Genomic Data Acquisition

All 459 whole fungal transcriptomes, as well as the corresponding RNA-sequencing expression data and eukaryotic cluster of orthologous (KOGs) annotations, were downloaded from the Joint Genome Institute's Mycocosm database (https://mycocosm.jgi.doe.gov). Only coding sequences (CDS) longer than 150bp with annotated start and stop codons were retained for downstream analysis.

**RNA sequencing***:* Most transcriptomes were sequenced using Illumina from RNA samples collected across multiple conditions and pooled together. From these samples libraries were prepared using tube- or plate-based RNA sample processing protocols. Plate-based RNA sample preparation was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Illumina's TruSeq Stranded mRNA HT Sample Prep Kit utilizing poly-A selection of mRNA following the manufacturer's guide starting with 1 µg total RNA per sample and 8-10 PCR cycles were used for library amplification. For the tube-based method, stranded cDNA libraries were generated using the Illumina Truseq Stranded mRNA Library Prep Kit, for which mRNA was purified from 1 µg (200 ng for low input RNA) of total RNA using magnetic beads containing poly-T oligos, fragmented, and reversed transcribed using random hexamers and SSII (Invitrogen) followed by second-strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and eight cycles (ten cycles for low input RNA) of PCR. The prepared libraries were quantified using the Next-generation Sequencing Library qPCR Kit (KAPA Biosystems) and run on a Roche LightCycler 480 real-time PCR instrument. Sequencing of the flowcells was performed on the Illumina HiSeq or NovaSeq, following a 2×150 indexed run recipe. Additional details may be available from the corresponding genome publications listed on MycoCosm (https://mycocosm.jgi.doe.gov/fungi/fungi.info.html).

**tRNA gene prediction**: tRNA genes were predicted with tRNA-scanSE2.0 (Lowe and Chan 2016) with eukaryotic-specific parameters. For quality control, only high-confidence tRNA genes with a covariance score of at least 50 were retained for analyses. tRNA gene copy number was used as the proxy for tRNA abundance.

*Seleno-cysteine tRNA identification*

High-confidence tRNA genes are assigned a tRNA-scanSE score of 50 and over. After applying the cut-off covariance score of 50, 6 genomes still retained high-scoring Sec-tRNA genes. To independently validate these tRNAscanSE predictions, these genomes were re-analyzed with another highly accurate but more conservative general-purpose tRNA gene finder aragorn1.2.38 (Laslett & Canback, 2004) using eukaryotic-specific parameters and a Sec-tRNA specific gene finder Secmarker (2015 Guigo). The final SeC positive species were taken as an overlap of any of these general gene-finders with the specialized Secmarker program.

### Codon Usage Metrics

The effective number of codons (ENC), which measures the degree of synonymous codon bias of gene or genome, was computed from coding sequences using CodonW 1.4.4 (Peden, 1995; Wright 1990).

The theoretical ENC is the expected value estimated solely based on GC3% due to neutral mutational bias. The theoretical ENC of a gene *g* that is only influenced by GC3-compositional bias was computed according to (Wright 1990) using custom python3 code.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

$$ENC_{theo} = 2 + GC3_g + (29 / GC3_g^2 + (1\text{-}GC3_g)^2)$$

**G+C composition at 3rd codon position (GC3-content):** GC3%was computed using CodonW 1.4.4 (Peden 1995)

**Relative synonymous codon usage (RSCU)** is the ratio of observed usage to the expected uniform usage within its amino acid class. RSCU is invariant to sequence length or amino acid composition. The RSCU of the 59 degenerate codons was computed using custom python3 scripts according to (Sharp and Li,1987). Six-fold amino acids (Leucine, Serine, Arginine) were split into 2-fold and 4-fold codon groups.

**Preferred and Non-preferred codons** were selected in each species based on the top and bottom 10% of expressed coding sequences. A codon is considered preferred if its RSCU value was significantly higher in the highly expressed CDS set (Mann-Whitney *U* test, Benjamini-Hochberg adjusted p-value <0.05). Conversely, a non-preferred codon reported a significantly higher RSCU in the low expressed CDS set. With this definition, more than one synonymous codon of an amino acid may be preferred or non-preferred.

**Frequency of optimal codons (Fop):** Fop was computed according to (Ikemura 1981) using custom python3 scripts based on optimal codons derived from a reference set of top 30 highly expressed ribosomal genes.

**Translation Bias Score, TBS:** We introduce our TBS to measure the extent to which the codon usage of an organism's gene expression reflects adaptation for the cellular tRNA supply based on the equation:

*((fraction of preferred codons decoded by major tRNAs) – (fraction of non-preferred codons decoded by major tRNAs)) / (sum of the fractions)*

Major tRNAs are the most abundant tRNA isoacceptor within an amino acid class, either based on gene copy number or tRNA concentration. In this paper, we used gene copy number as a proxy for cellular tRNA concentrations.

## Comparative Phylogenetic Calculations

We downloaded from the fungal phylogenetic tree from Joint Genome Institute Mycocosm (https://mycocosm.jgi.doe.gov). The phylogenetic tree was then pruned using Dendropy package (v4.40) in python3.7. Taxonomic ranks were obtained from National Center for Biotechnology Information (NCBI).

Phylogenetic Independent Contrasts (PIC) models the trait covariation according to the formula $Y = \beta X + \varepsilon$, where Y and X are traits and β is the evolutionary correlation coefficient that quantifies the degree of coevolution between traits X and Y. Phylogenetic independent contrast was computed with the picante package in R (Kembel et al, 2010).

Maximum likelihood continuous ancestral trait reconstruction was performed using contMap(model=Brownian motion) function from the package phytools (Revell 2011). Blomberg's K statistic were computed using the phylosig() function in phytools library (Revell 2012) implementation in R which p-values are calculated based on 100 permutations.

Fitting and evaluation of maximum likelihood macroevolutionary models for continuous character evolution were performed using the geiger library (v2.0.6.3; Pennell et al, 2014) in R.

## Correlation, Hypothesis and Multivariate analyses

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

All correlation and significance tests were performed using the scipy (v1.3.1) and statsmodels(0.10.1) libraries in python3.7. Principal component analysis (PCA) was performed using the scikit-learn (v0.21.3; Pedregrosa et al 2011) in python3.7.

## Supervised Neural Network Codon2Vec

**Codon2Vec** is an artificial neural network (ANN) that learns the species-specific dependency between the codon composition (features) of a coding sequence (CDS) and expression level. A neural network is a class of machine learning algorithms that uses layers of interconnected computation nodes to learn complex patterns that underlie the data. We implemented and trained Codon2Vec using the keras (v2.2.4) (Chollet 2015) with tensorflow v1.8 backend, and scikit-learn libraries in Python3.7.

Available as a command-line tool for download at this github repository.

*Dataset Collection and Preprocessing*

We selected CDS from the top and bottom 10% of expression distribution relative to the mean expression. Each CDS was represented as a vector of codons in sequence. Then each unique codon is assigned a unique integer (*tokenization)* such that each CDS becomes recoded as a vector of integers. Finally, the lengths of the CDS were set to a fixed size of 2000, either by trimming longer sequences or padding with zeroes. The input and output data were shuffled and partitioned into 70% training set, 20% validation set and 10% test set. The training set is used to learn the model weights, whereas the validation set is used to fine-tune the model's generalizability by evaluating whether the model is over- or under-fitting on data it was not trained on. The final evaluation is performed on the test set.

*Model training*

Codon2Vec is a feedforward ANN with 3 fully connected computation layers - embedding layer, rectified linear unit (ReLu) activation layer and sigmoid output layer. We also incorporated "drop-out" regularization to reduce overfitting. Each layer is described in more detail in the subsequent paragraphs.

*Learning optimized model weights*

A node is the fundamental computation unit of an ANN. In a fully connected ANN, all nodes of a layer receive the *weighted* output of each node from the previous layer. The weights (**W**) represent the relative importance of a node to the model performance. During the forward pass of training, the input (**X**) undergoes a series of matrix multiplications and non-linear transformations ($\phi$) as it flows sequentially between nodes in each layer until the predicted output is generated in the final layer.

Generating prediction: $\hat{Y} = \phi(W^T X)$ , where X and W are matrices

Model weights (**W)** were randomly initialized based on the Glorot uniform distribution. We chose the binary cross-entropy loss as the optimization objective that computes the error between the predicted output ($\hat{Y}$) and ground truth (**Y**).

$$\textbf{Loss}(Y, \hat{Y}) = -(Y * \log(\hat{Y}) + (1 - Y) * \log(1 - \hat{Y}))$$

where $(\hat{Y}) \in \{0,1\}$ and Y is binary encoded as 0 or 1

In the backward pass of training, the contribution of the current set of weights (W) to the model error is computed by taking the partial derivative of the loss function with respect to

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

each layer's weights (W). The weights are then updated by their gradients in the direction that minimizes the loss function. Weights were tuned via backpropagation using the Adam optimization, a variant of stochastic gradient descent based on adaptive learning (Kingma et al 2014).

*Model architecture*

The first layer serves as feature extraction by learning the weights that map each of the 64 unique codon features to a meaningful dense real-valued 4-dimensional vector (embeddings) in Euclidean space. The number 4 is a hyperparameter. The advantages of embedding representation are: 1) the features that maximize model performance are learned directly from the CDS 2) the numerical transformation of codons makes modeling amenable to neural networks, and 2) it is more computationally efficient than the alternative one-hot representation as each codon would have been assigned a 1x64 dimensional sparse vector of mostly zeroes compared to Codon2Vec's 1x4 dimensional dense vector.

The second layer applies the ReLu activation function to the weighted sum of outputs from the embedding layer. The ReLu function is a widely preferred non-linear transformation for the inner (hidden) layers of neural networks because it speeds up convergence, and is robust to vanishing gradients.

$$\textbf{\textit{ReLu (x) = max(0, x)}}$$

Finally, the output layer applies the sigmoid function that maps the continuous values to a real value between 0 and 1, such that the final output is a 2-dimensional vector of the prediction probabilities for each expression class.

*Model Evaluation*

We evaluated Codon2Vec's predictive performance using misclassification error, sensitivity, specificity and precision on the test set. Let TP, TN, FP, FN denote true positives, true negatives, false positives and false negatives, respectively:

$$\text{Misclassification error} = 1 - (TP+TN) / (TP+TN+FP+FN)$$
$$\text{Sensitivity} = TP/(TP+FN)$$
$$\text{Specificity} = TN / (TN+FP)$$
$$\text{Precision} = TP/ (TP+FP)$$

Furthermore, we plotted the receiver-operating characteristic curves and calculated the area under the ROC curve (AUC-ROC).

**Data availability**

All derived data files are included in the supplemental information. Custom python3 and R scripts used in analysis are available on this github repository.

**Acknowledgments**

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

We would like to thank Cathie Aime, Manuel Alfaro, Mikael Anderson, Ólafur Andrésson, Betsy Arnold, Scott Baker, Greg Bonito, Tom Bruns, Kathryn Bushley, Nadya Cardona, Ying Chang, In-Geol Choi, Laurie Connell, Luis Corrochano, Pedro Crous, Gunther Doehlemann, Sebastien Duplessis, Paul Dyer, Dan Eastwood, Romina Gazis-Seregina, John Gladden, Dave Greenshields, Andrii Gryganskyi, Milan Gryndler, Richard Hamelin,

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

Coleen Hansel, Patrik Inderbitzin, Tim James, Havard Kauserud, Konstantin Krutovsky, Dan Lindner, Jon Magnuson, Francis Martin, Sundy Maurice, Laszlo Nagy, Don Natvig, Robin Ohm, Kabir Peay, Amy Powel, Daniel Raudabaugh, Steven Singer, Joey Spatafora, Jason Stajich, Rytas Vilgalys, Todd Ward, and Silke Werth for sharing their not published data produced by JGI for several JGI CSP projects and allowing use of the pre-publication data in this study.

**References**

1. Ahrendt SR, Quandt CA, Ciobanu D, Clum A, Salamo, A, Andreopoulos B, Cheng JF, Woyke T, Pelin A, Henrissat B, Reynolds, NK, Benny GL, Smith ME, James TY, Grigoriev IV, et al. 2018. Leveraging single-cell genomics to expand the fungal tree of life. *Nat Microbiol.* 3:1417–1428.
2. Badet T, Peyraud R, Mbengue M, Navaud O, Derbyshire M, Oliver RP, Barbacci A, Raffaele, S. 2017. Codon optimization underpins generalist parasitism in fungi. *eLife*, 6:e22472. 2
3. Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*. 66: 2369–2383.
4. Berbee ML, James T Y, Strullu-Derrien C. 2017. Early diverging fungi: diversity and impact at the dawn of terrestrial life. *Ann. Rev. Microbiol*. 71:41–60.
5. Blomberg SP, Garland T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution.* 57:717–745.
6. Botzman M, Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol*. 1210: R109.
7. Bour T, Mahmoudi N, Kapps D, Thiberge S, Bargieri D, Ménard R, Frugier M. Apicomplexa-specific tRip facilitates import of exogenous tRNAs into malaria parasites. 2016. *Proc Natl Acad Sci USA*. 11317:4717-4722.
8. Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*. 1293: 897–907.
9. Butler MA, King AA. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am Nat*. 2004;164: 683–695.
10. Butler MA and King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.* 164:683–695.
11. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* 1412:355–367.
12. Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* 44D1: D184–D189.
13. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA.* 101:3480–3485.
14. Chollet F. 2015. keras. *GitHub*. https://github.com/fchollet/keras
15. Disconis P, Goel S, Holmes S. 2008. Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat*. 2:777– 807.
16. dos Reis M, Wernisch L. 2004. Estimating translational selection in eukaryotic genomes. *Mol Biol Evol*. 262: 451–461.
17. Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*. 167:287–289.
18. Felsenstein, J. 1985 Phylogenies and the comparative method. *Am Nat.* 125, 1–15
19. Field KJ, Rimington WR, Bidartondo MI, Allinson KE, Beerling DJ, Cameron DD, Duckett JG, Leake JR, Pressel S. 2015. First evidence of mutualism between ancient plant lineages Haplomitriopsida liverworts and Mucoromycotina fungi and its response to simulated Palaeozoic changes in atmospheric CO2. *New Phytol*. 205: 743-756.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

20. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. 2018.Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci USA.* 1521: E4940-E4949.
21. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. 2016. Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell*. 166(3):679-690.
22. Garland Jr T, Harvey PH, Ives AR. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* 41:18-32.
23. Gaya E, Fernández-Brime S, Vargas R, Lachlan RF, Gueidan C, Ramírez-Mejía M, Lutzoni F. 2015. The adaptive radiation of lichen-forming Teloschistaceae is associated with sunscreening pigments and a bark-to-rock substrate shift. *Proc Natl Acad Sci U S A.* 112(37):11600-11605.
24. Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol*. 7:481.
25. Grantham R, Gautier C, Gouy M, Mercier R, Pav A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acid Res*. 81: r49–r62.
26. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. 2014. MycoCosm portal: Gearing up for 1000 Fungal Genomes. *Nuc Acids Res*. 421: D699-D704.
27. Harismendy O, Gendrel CG, Soularue P, Gidrol X, Sentenac A, Werner M, Lefebvre O. 2003. Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J.* 22(18):4738-4747.
28. Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet*. 42:287–299.
29. Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol. 2511:2279–2291.
30. Hoernes TP, Faserl K, Juen MA, Kremser J, Gasser C, Fuchs E, Shi X, Siewert A, Lindner H, Kreutz C, Micura R, Joseph S, Höbartner C, Westhof E, Hüttenhofer A, Erlacher MD. 2018. Translation of non-standard codon nucleotides reveals minimal requirements for codon-anticodon interactions. *Nat Comm.* 9:4865.

31. Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol*. 1584:573-597.
32. James T, Kauff F, Schoch C., *et al.* 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature.* 443, 818–822.
33. Janbon G, Quintin J, Lanternier F, d'Enfert C. 2019. Studying fungal pathogens of humans and fungal infections: fungal diversity and diversity of approaches. *Genes Immun.* 20: 403–414.
34. Kamilar JM, Cooper N. 2013. Phylogenetic signal in primate behaviour, ecology and life history. *Phil Trans R Soc B*. 368: 20120341.
35. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*. 26:1463–1464.
36. LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas. A 2019. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet*.157: e1008304.
37. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*. 321:111-116.
38. Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN. 2007. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol*. 8:R198
39. Losos JB. 2011. Convergence, adaptation, and constraint. *Evolution*. 65, 1827-1840.
40. Losos, J.B. 2011. Seeing the forest for the trees: the limitations of phylogenies in comparative biology: American Society of Naturalists Address. *Am. Nat*. 177:709–727.
41. Mao H and Wang, H. 2019. Resolution of deep divergence of club fungi phylum Basidiomycota. *Synth. Syst. Biotechnol*. 4:225–231.


Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

42. Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 196:330–338.
43. Marck C, Grosjean H. 2002. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*. 810:1189–1232.
44. Mariotti M, Salinas G, Gabaldón T, Gladyshev VN. Utilization of selenocysteine in early-branching fungal phyla*.* 2019. Utilization of selenocysteine in early-branching fungal phyla. *Nat Microbiol.* 4:759–765.
45. McEntee JP, Tobias JA, Sheard C, Burleigh JG. 2018. Tempo and timing of ecological trait divergence in bird speciation*. Nat Ecol Evol*. 2:1120–1127.
46. Nagy LG, Házi J, Szappanos B, Kocsubé S, Bálint B, Rákhely G, Vágvölgyi C, Papp T. 2012. The evolution of defense mechanisms correlate with the explosive diversification of autodigesting Coprinellus mushrooms Agaricales Fungi. *Syst Biol* .61:595–607
47. Naranjo-Ortiz MA, Gabaldon T. 2019. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biol Rev Camb Philos Soc.* 94: 2101-2137.
48. Nguyen LH and Holmes, S. 2019. Ten quick tips for effective dimensionality reduction. *PLoS computational biology*. 15(6): e1006907.
49. Nishikura, K. 2016. A-to-I editing of coding and non-coding RNAs by ADARs. Nat. Rev. *Mol. Cell Biol.* 17, 83–96.
50. Novoa EM, Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A role for tRNA modifications in genome structure and codon usage. *Cell.* 1491:202-213.
51. Novoa EM, Jungreis I, Jaillon O, Kellis M. 2019. Elucidation of Codon Usage Signatures across the Domains of Life. *Mol Biol Evol.* 3610:2328-2339.
52. Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*. 401:877–884.
53. Palidwor GA, Perkins TJ, Xia X. 2010. A general model of codon bias due to GC mutational bias. *PLoS One*. 5:13431.
54. Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol*. 20(2):237-243.
55. Peden JF. 1999. Analysis of codon usage. PhD Thesis, University of Nottingham.
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: Machine Learning in Python. *JMLR*; 12: 2825-2830.
57. Pennell M and Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci*. 1289:90–105.
58. Pennell MW, Eastman JM Slater GJ, Brown JW, Uyeda JC, Fitzjohn RG, Harmon LJ. 2014. Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*15, 2216–2218.
59. Pichereira-Donoso D, Harvey LP, Ruta M. 2015. What defines an adaptive radiation? Macroevolutionary diversification dynamics of an exceptionally species-rich continental lizard radiation. *BMC Evol Biol*. 15:153.
60. Powell JR, Moriyama E. 1997. Evolution of codon usage in *Drosophila*. *Proc Natl Acad Sci USA.* 94:7784–7790.
61. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. 2015. Codon optimality is a major determinant of mRNA stability. *Cell*. 160:1111–1124.
62. R Development Core Team. 2011. R: A Language and Environment for Statistical Computing.
63. Rafels-Ybern A, Torres AG, Grau-Bove X, Ruiz-Trillo I, Ribas de Pouplana L. 2018. Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla. *RNA Biology*. 15(4-5):500-507.
64. Rak R, Dahan O, Pilpel Y. 2018. The Couplers of Genomics and Proteomics. *Annu Rev Cell Dev Biol*. 34:239-264.


Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

65. Revell LJ, Harmon LJ, Collar DC. 2008. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol*. 57, 591–601.
66. Revell LJ. Two new graphical methods for mapping trait evolution on phylogenies. 2013. *Methods Ecol Evol*. 4:754–9.
67. Revell, L.J. 2012 Phytools: phylogenetic tools for comparative biology and other things. *Methods in Ecology and Evolution*. 3: 217-223.
68. Rogalski M, Karcher D, Bock R. 2008. Superwobbling facilitates translation with reduced tRNA sets*. Nat. Struct. Mol. Biol*.; 15:192-198
69. Roller M, Lucić V, Nagy I, Perica T, Vlahovicek K. 2013. Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res*. 4119:8842–8852.
70. Sánchez-García M and Matheny PB. 2017. Is the switch to an ectomycorrhizal state an evolutionary key innovation in mushroom-forming fungi? A case study in the Tricholomatineae Agaricales. *Evolution*. 711:51-65.
71. Santesmasses D, Mariotti M, Guigó R. 2017. Computational identification of the selenocysteine tRNA tRNASec in genomes. *PLoS Comput Biol*. 132:e1005383.
72. Sepala S, Wilken E, Knop D, Solomon K, O'Malley M. 2017. The importance of sourcing enzymes from non-conventional fungi for metabolic engineering and biomass breakdown. *Metabolic Engineering*. 44:45-59.
73. Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*. 3651544:1203–1212.
74. Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*. 1413:5125–5143.
75. Simpson GG. 1953. The major features of evolution. New York: Columbia University Press.
76. Smith SD. 2010. Using phylogenetics to detect pollinator-mediated floral evolution. *New Phytologist*. 188: 354–363.
77. Spatafora JW, Aime MC, Grigoriev IV, Martin F, Stajich JE, Blackwell M. 2017. The fungal tree of life: from molecular systematics to genome-scale phylogenies. *Microbiol Spectr* 5: FUNK-0053-2016
78. Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW. 2009. The fungi. *Curr Biol*. 1918: R840-R845.
79. Supek F and Vlahoviček, K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*. 11:463.
80. Toome M, Ohm RA, Riley RW, James TY, Lazarus KL, Henrissat B, Albu S, Boyd A, Chow J, Clum A, Heller G, Lipzen A, Nolan M, Sandor L, Zvenigorodsky N, Grigoriev IV, Spatafora JW, Aime MC. 2014. Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen Mixia osmundae. *New Phyt.* 202(2):554-564.
81. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*. 1412:344–354.
82. Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, Andreopoulos W, Angelini C, Antonín V, Barry KW, *et al.* 2019. Megaphylogeny resolves global patterns of mushroom evolution. *Nat Ecol Evol.* 3:668–678.
83. Velandia-Huerto CA, Berkemer SJ, Hoffmann A, Retzlaff N, Romero Marroquín LC, Hernández-Rosales M, Stadler PF, Bermúdez-Santana CI. 2016. Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies. *BMC Genomics*. 17(1):617.
84. Wright F. 1990 The 'effective number of codons' used in a gene. *Gene.* 871: 23–29.
85. Yang KK, Wu Z, Bedbrook CN, Arnold FH. 2018. Learned protein embeddings for machine learning. *Bioinformatics*. 1:7.
86. Yannai A, Katz S, Hershberg R. 2018. The codon usage of lowly expressed genes is subject to natural selection. *Genome Biol Evol*. 105: 1237–1246.

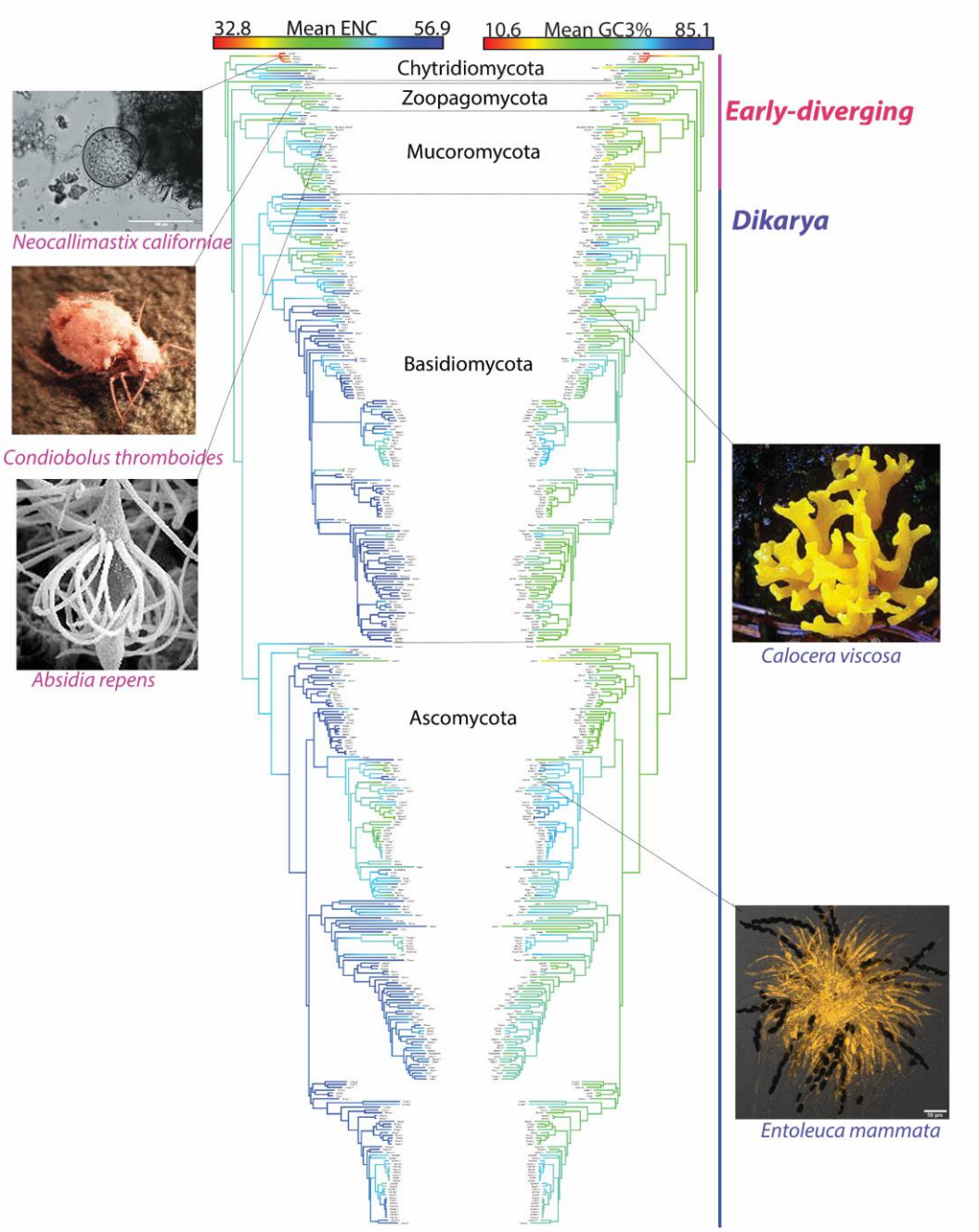

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

87. Yona AH, Bloom-Ackermann Z, Frumkin I, Hanson-Smith V, Charpak- Amikam Y, Feng Q, Boeke JD, Dahan O, Pilpel Y. 2013. tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* 2: e01339.
88. Zhao F, Zhou Z, Dang Y, Na H, Adam C, Lipzen A, Ng V, Grigoriev IV, Liu Y. 2021. Genome-wide role of codon usage on transcription and identification of potential regulators. *Proc Natl Acad Sci U S A.* 1186: e2022590118.
89. Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol.* 267:1571–1580.
90. Zhou Z, Dang Y, Zhou M, Li L, Yu C, Fu J, Chen S, Liu Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl Acad. Sci. USA.* 113: E6117-E6125.
91. Zouridis H. and Hatzimanikatis V. 2008. Effects of codon distributions and tRNA competition on protein translation*. Biophys. J.* 95:1018–1033.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

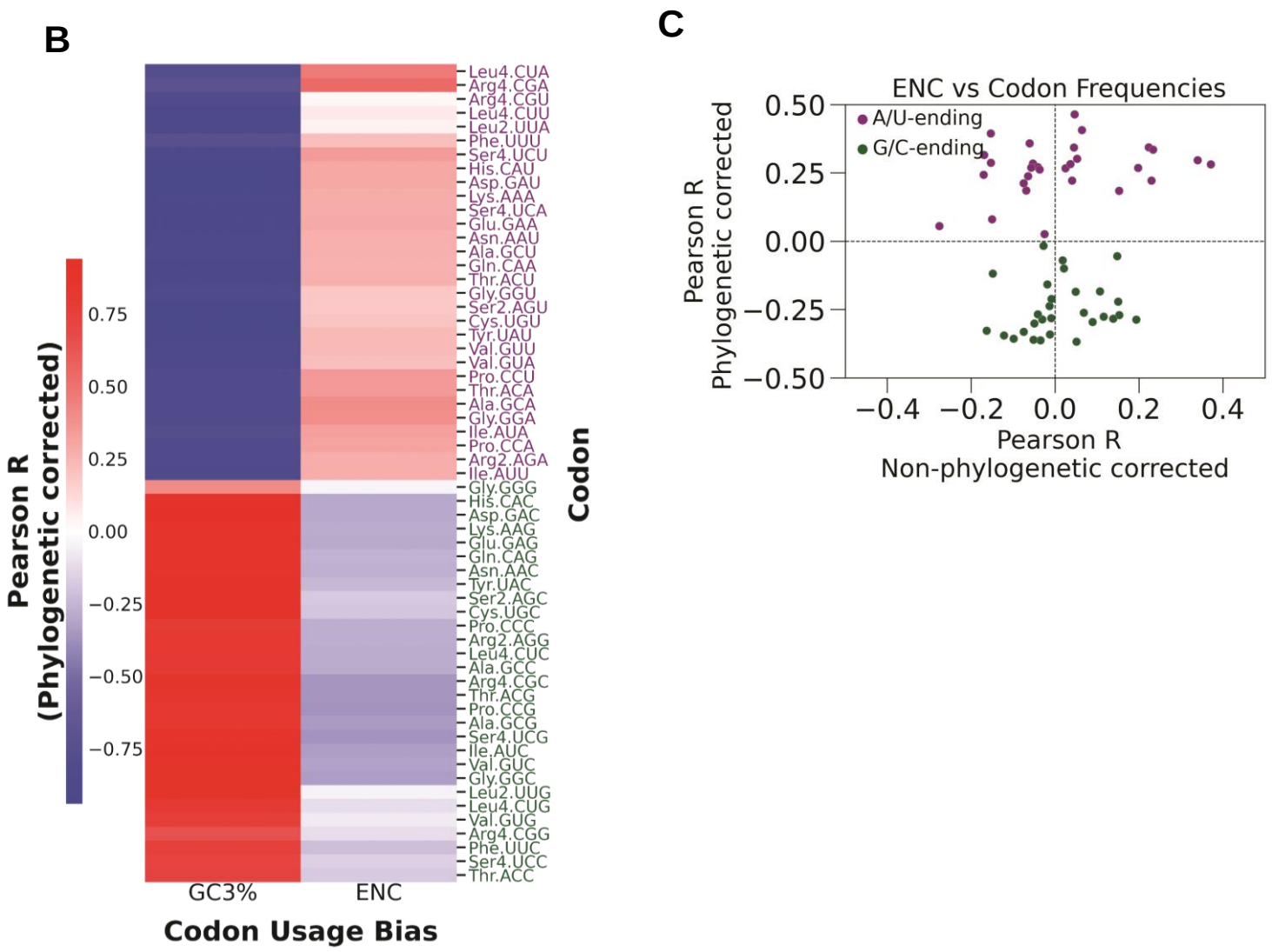
# Figure 1

**A**



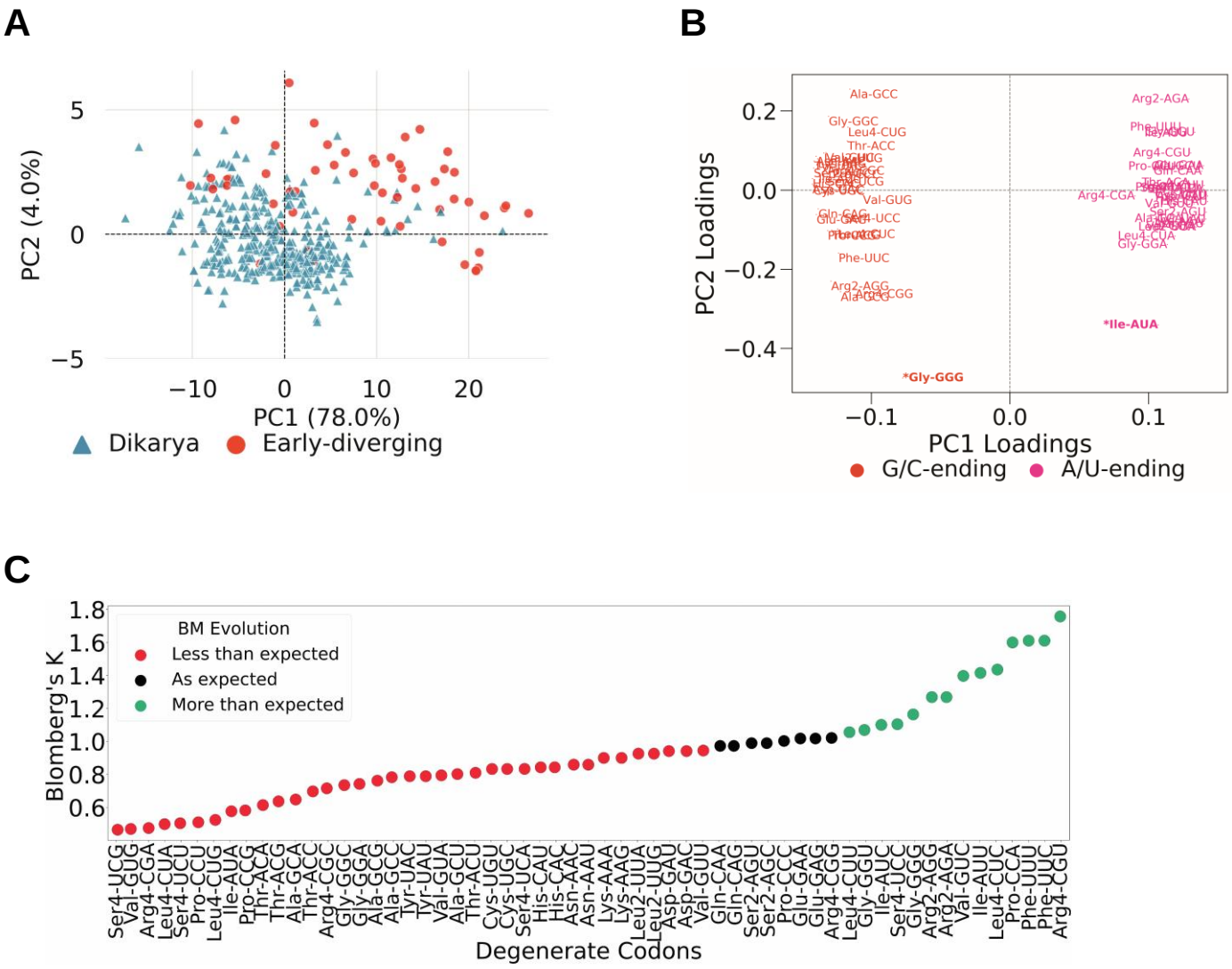

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

# Figure 1



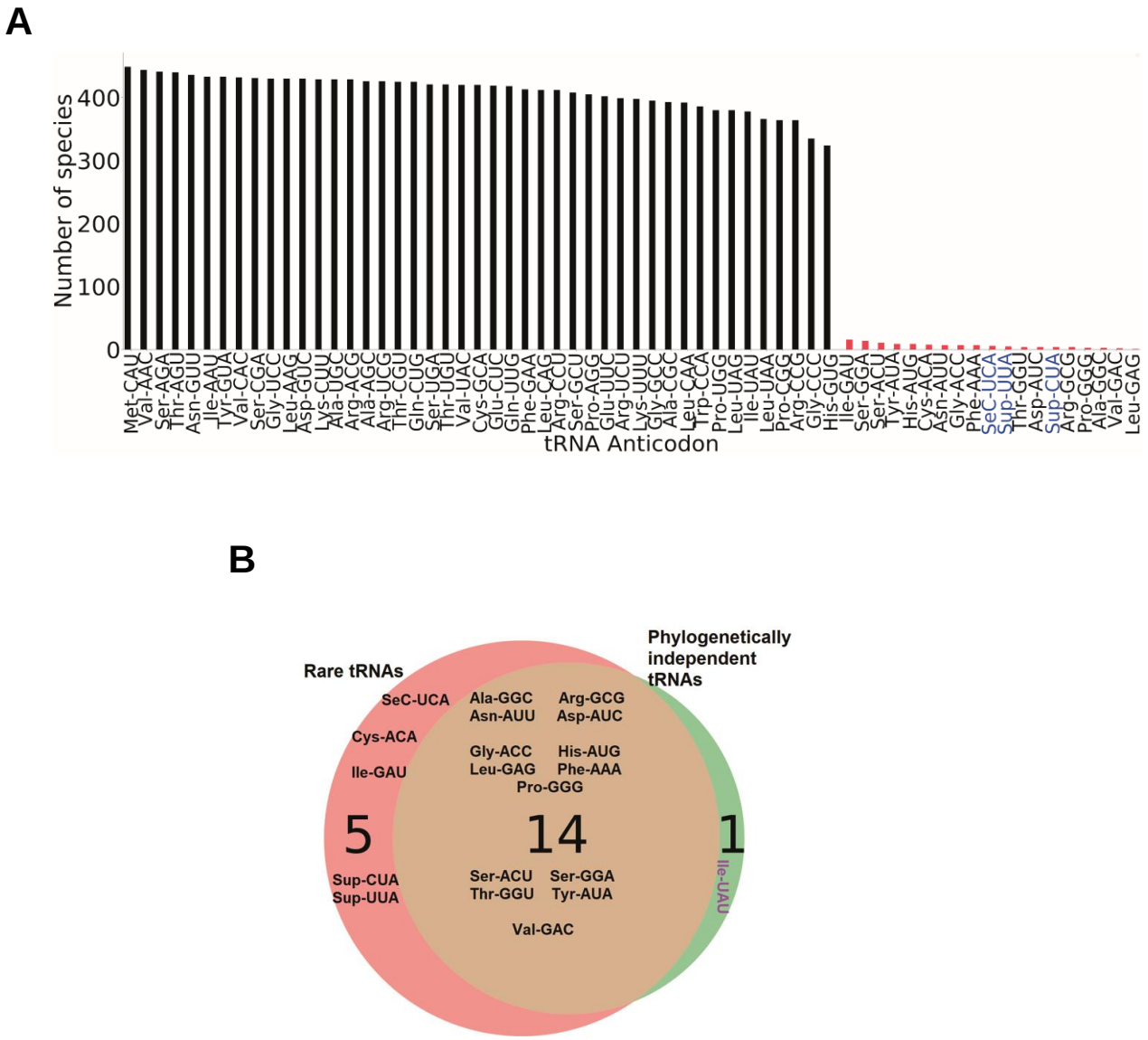

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

# Figure 2



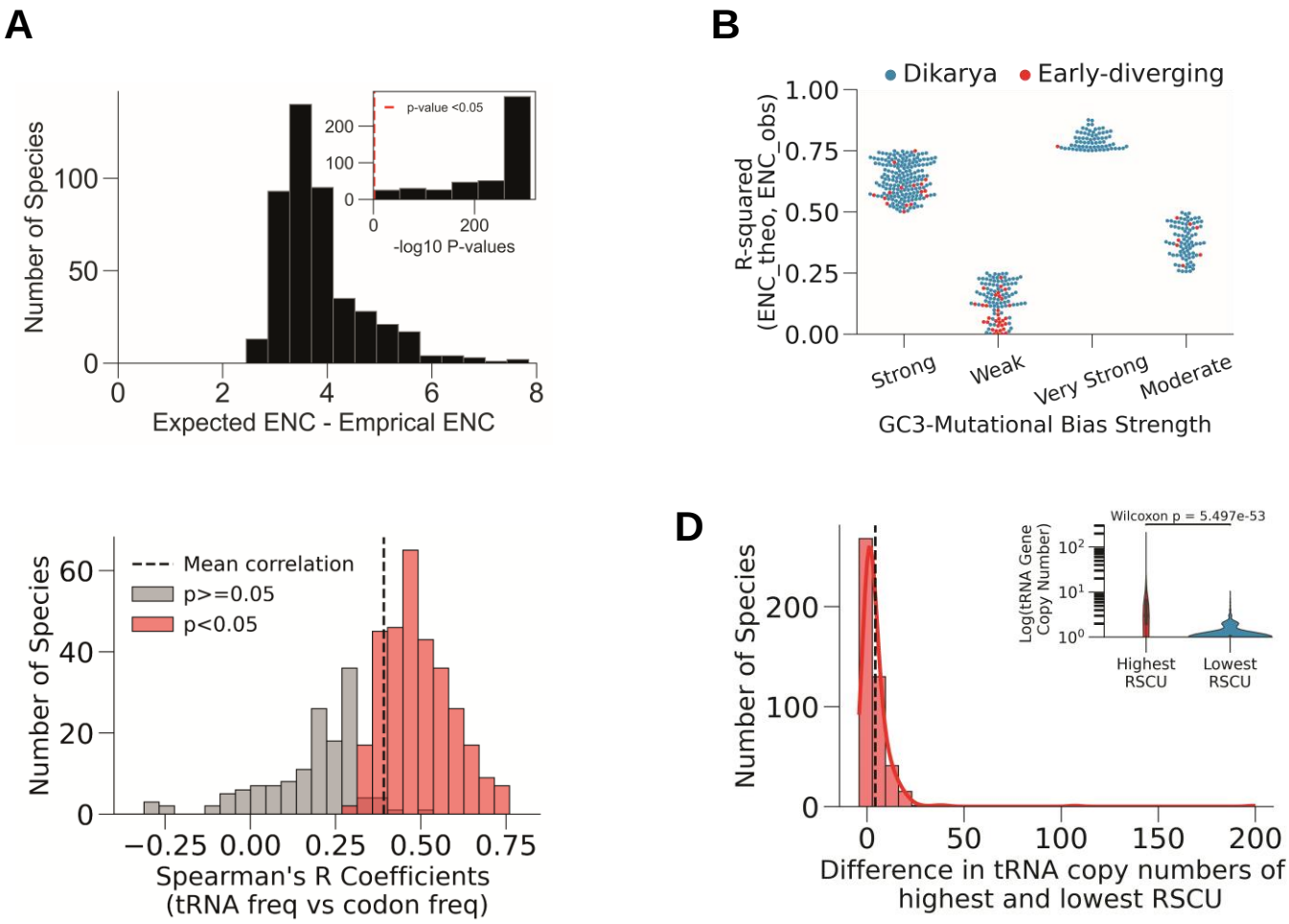

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

# Figure 3



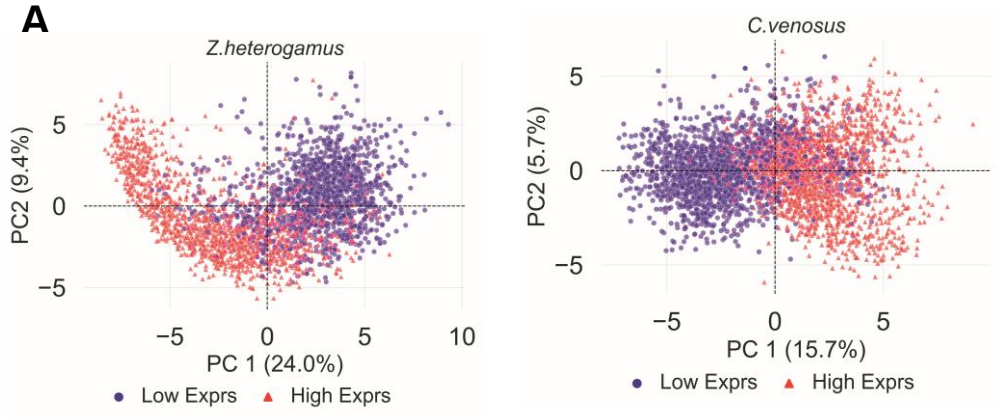

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

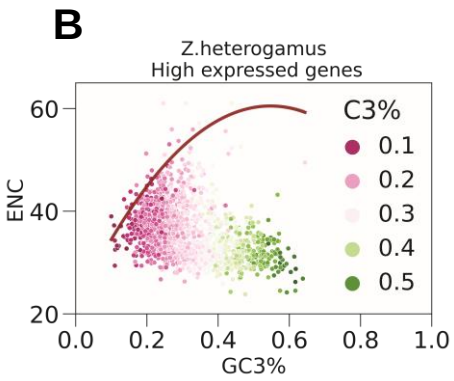
# Figure 4

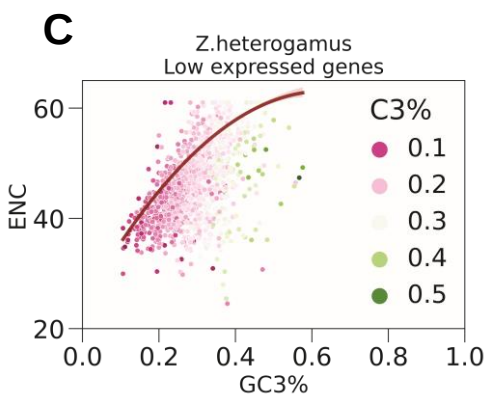


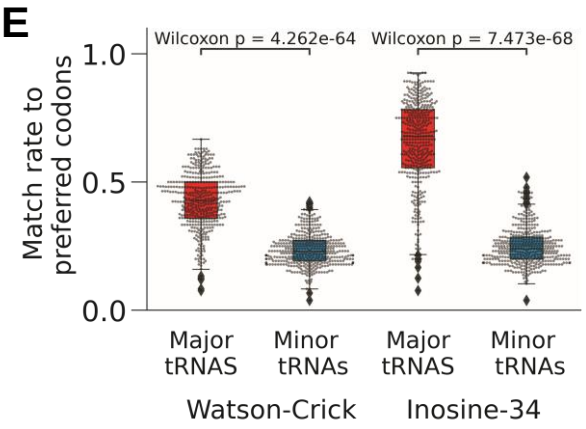

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

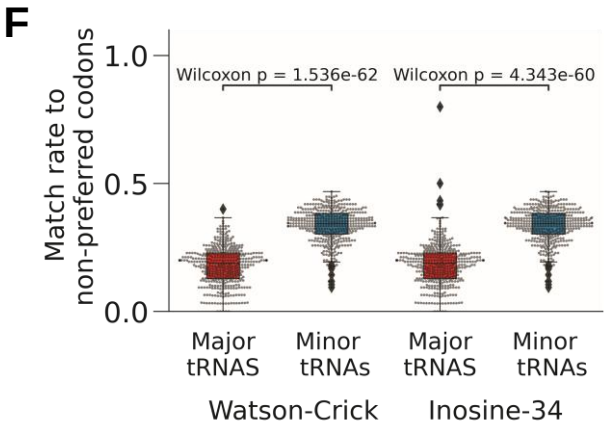
# Figure 5

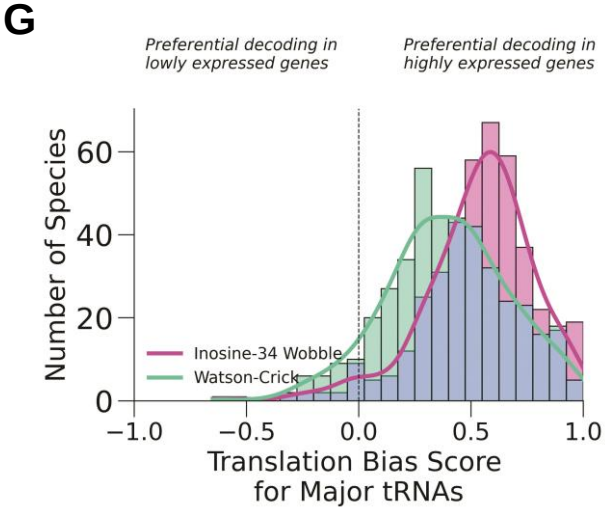


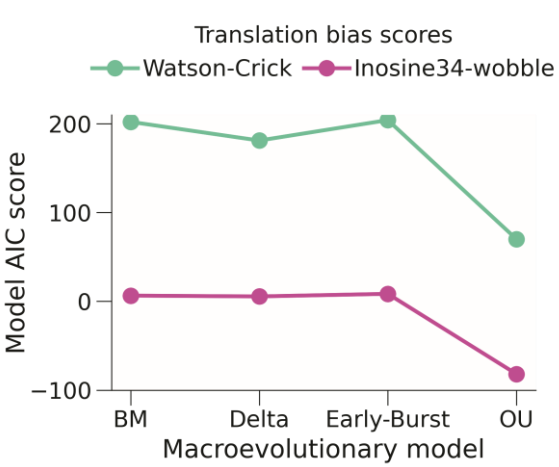





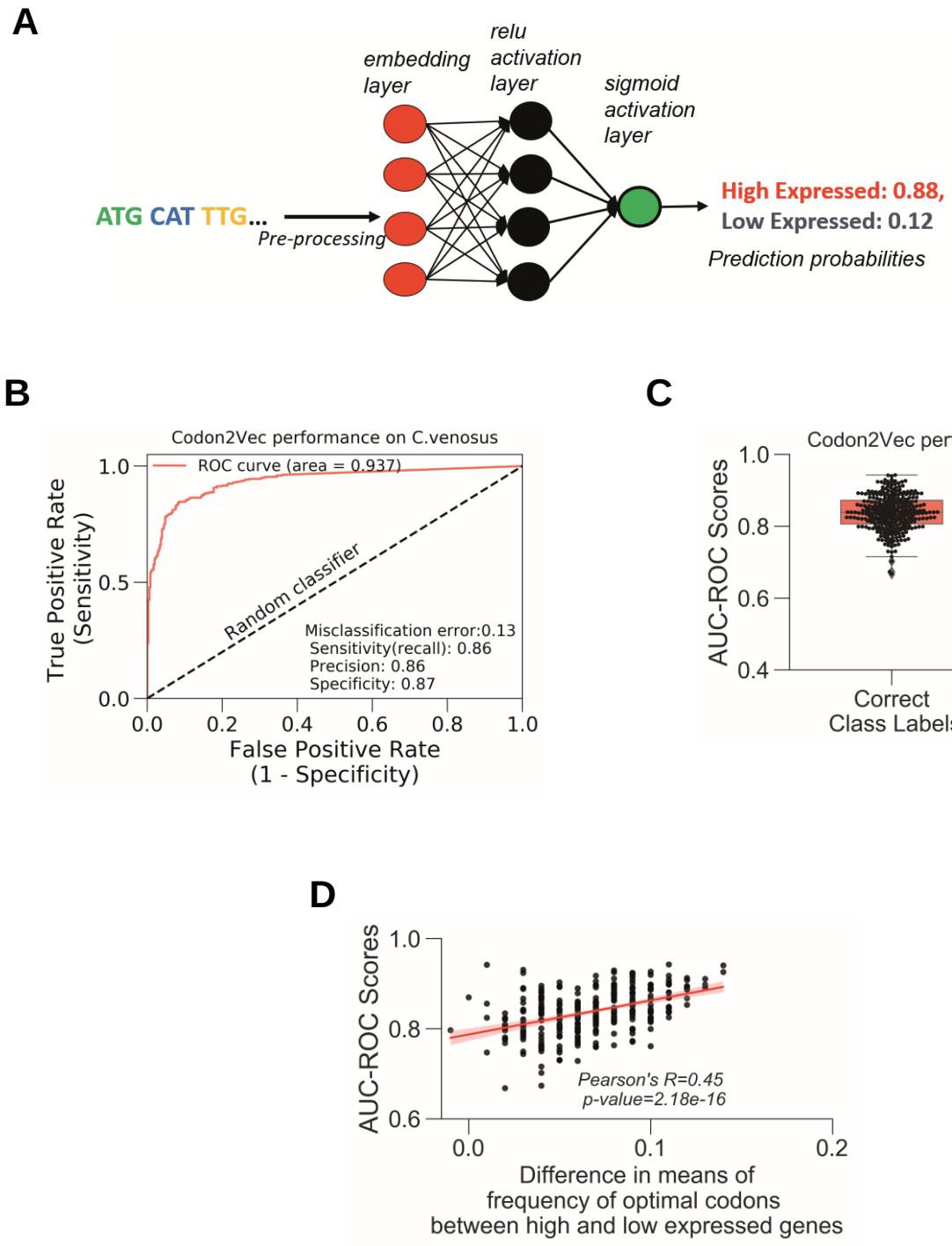




Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022

# Figure 6




Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab372/6513383 by guest on 25 January 2022